3D Human Pose from Silhouettes by Relevance Vector Regression

Ankur Agarwal and Bill Triggs
GRAVIR-INRIA-CNRS, Grenoble, France http://lear.inrialpes.fr {Ankur.Agarwal,Bill.Triggs}@inrialpes.fr

1 In Brief

Goal

- Recover ***3D human body pose*** from monocular ***image silhouettes***
  - 3D pose = joint angles
  - use either individual images or video sequences

Contributions

- "Model-free" learning based approach
  - no explicit 3D model — recovers 3D pose by direct regression against robust silhouette descriptors
- Sparse kernel regressor trained using human motion capture data
- Regression based filtering for resolving reconstruction ambiguities
- Mean errors of only 4-6° per joint angle on test sequences

Method

Silhouette → Descriptor vector → RVM regressor → 3D pose vector

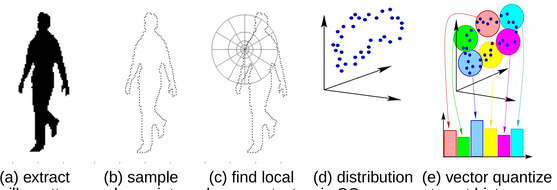
2 Silhouette Descriptors

Why Silhouettes

\+
- Relatively simple and low-level
- Capture most of the available pose information
- Insensitive to surface attributes (clothing colour, texture..)

\−
- Frequently distorted by background subtraction / shadows
- Ambiguity: internal details and depth ordering are hidden

Robust encoding of local shape — Shape Context Histograms



(a) extract silhouette (b) sample edge points (c) find local shape contexts (d) distribution in SC space (e) vector quantize to get histogram

3 Training and Test Data

- For the movements, we use real human motion capture data — captures ***typical*** human movements, not just *possible* ones
- Synthesize silhouettes with POSER human modeller (Curious Labs) — somewhat artificial, but gives ground truth for testing, allows a wide range of training viewpoints.
- Also tested on real sequences of other people (without ground truth)

4 Nonlinear Regression Model

Given input: shape context histogram vector $\mathbf{x}$
Desired output: 3D human pose vector $\mathbf{y}$

$$\mathbf{y} \;=\; \mathbf{A}\,\mathbf{f}(\mathbf{x}) + \boldsymbol{\epsilon} \;\equiv\; \sum_{k=1}^{p} \mathbf{a}_k\,\phi_k(\mathbf{x}) + \boldsymbol{\epsilon}$$

- $\mathbf{f}(\mathbf{x}) = \big(\phi_1(\mathbf{x}) \;\cdots\; \phi_p(\mathbf{x})\big)^\top$: vector of scalar basis functions $\phi_k(\mathbf{x})$
- $\mathbf{A} \equiv \big(\mathbf{a}_1 \;\; \mathbf{a}_2 \;\cdots\; \mathbf{a}_p\big)$: matrix of weight vectors $\mathbf{a}_k$ ***to be learned***
- $\boldsymbol{\epsilon}$: residual error vector

4.1 Penalized Least Squares

Estimate $\mathbf{A}$, given a set of training pairs $\{(\mathbf{y}_i, \mathbf{x}_i) \,|\, i = 1 \ldots n\}$:

$$\mathbf{A} \;:=\; \arg\min_{\mathbf{A}} \left\{ \sum_{i=1}^{n} \|\mathbf{A}\,\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i\|^2 \;+\; R(\mathbf{A}) \right\}$$

- $\mathbf{x}_i$ enter only via ***feature vectors*** $\mathbf{f}_k(\mathbf{x}_i) \;=\; \phi_k(\mathbf{x}_i)$.
- $R(\mathbf{A})$ is a regularizer on $\mathbf{A}$ to control overfitting

Ridge Regression: $R(\mathbf{A}) = \lambda\|\mathbf{A}\|^2$

4.2 The Relevance Vector Machine

- A Bayesian-motivated approach to regression and classification
- Uses a singular ***power-law prior*** to aggressively prune unneeded weights, giving ***sparse solutions*** $\mathbf{A}$.
- Regularizer:

$$R(\mathbf{A}) \;=\; \nu \sum_{a} \log \|a\|$$

  - ν is the pruning / shrinkage strength
  - a can be the components, the columns, or the rows of $\mathbf{A}$

RVM Training Algorithm

0. Initialize $\mathbf{A}$ with ridge regression. Initialize the running scale estimates $a_{\text{scale}} = \|\mathbf{a}\|$ for the components or vectors $\mathbf{a}$.
1. Approximate the $\nu \log \|\mathbf{a}\|$ penalty terms with "quadratic bridges" $\nu\,(\mathbf{a}/a_{\text{scale}})^2 + \text{const}$ (the gradients match at a_{scale});
2. Solve the resulting linear least squares problem in $\mathbf{A}$;
3. Remove any components $\mathbf{a}$ that have become zero, update the scale estimates $a_{\text{scale}} = \|\mathbf{a}\|$, and continue from 1 until convergence.

Linear bases

- $\mathbf{f}(\mathbf{x}) \equiv \mathbf{x} \Rightarrow$ the RVM selects relevant silhouette *features*.

left arm right leg ——— torso and neck ———

Kernel bases

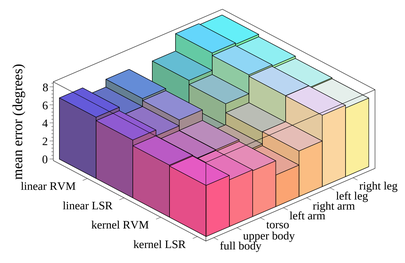
- $\mathbf{f}(\mathbf{x}) \;=\; (K(\mathbf{x}, \mathbf{x}_1) \;\cdots\; K(\mathbf{x}, \mathbf{x}_n))^\top$ where $K(\mathbf{x}, \mathbf{x}_i)$ is a kernel function instantiated at training examples $\mathbf{x}_i$
- RVM selects relevant training *examples*, here only 6%

5 Pose from Static Images

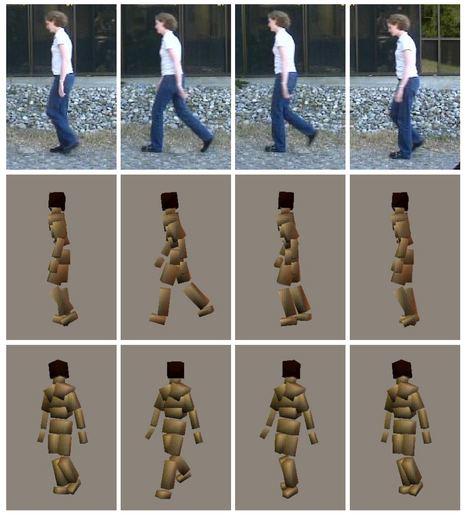
Reconstruction on test sequence

Spiral walking sequence not included in training data. Gaussian kernel regressor. Mean angular error per d.o.f = 6.0°

Summary of Regressors' Performance



Sample reconstructions from real images



Original (middle) and new (bottom) viewpoints

6 Pose from Video Sequences

Tracking Framework

- Tracking reduces glitches caused by silhouette ambiguities
- ***Regression based filtering*** for dynamical prediction and observation update ($\mathbf{x}$: 3D pose state, $\mathbf{z}$: Silhouette descriptor)

Dynamics

- A second order global dynamical model suffices:

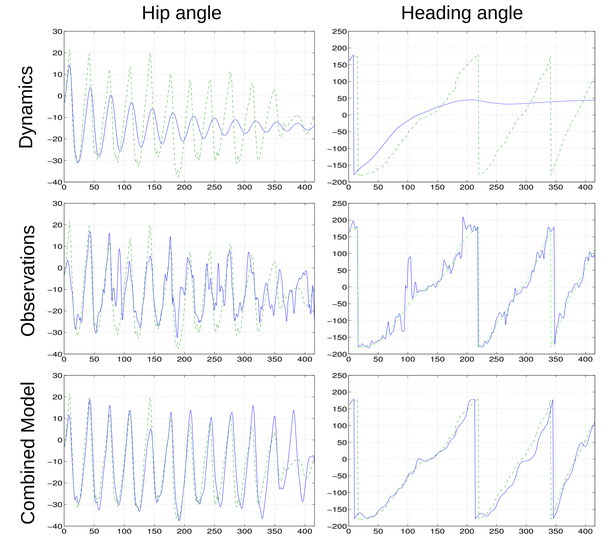
$$\check{\mathbf{x}}_t \;\equiv\; (\mathbf{I} + \mathbf{A})(2\mathbf{x}_{t-1} - \mathbf{x}_{t-2}) + \mathbf{B}\,\mathbf{x}_{t-1}$$

State-Sensitive Observation Update

- Nonlinear kernel regressor "selects" observation update to apply using state prediction
- Our full regression model also includes an explicit $\check{\mathbf{x}}_t$ term to represent the direct contribution of the dynamics

$$\hat{\mathbf{x}}_t \;\equiv\; \mathbf{C}\,\check{\mathbf{x}}_t + \sum_{k=1}^{p} \mathbf{d}_k\,\phi_k(\check{\mathbf{x}}_t, \mathbf{z}_t) \;=\; (\mathbf{C}\;\mathbf{D}) \begin{pmatrix} \check{\mathbf{x}}_t \\ \mathbf{f}(\check{\mathbf{x}}_t, \mathbf{z}_t) \end{pmatrix}$$

Results



Angles (in degrees) vs. Time. - - - Ground truth —— Estimate

7 Conclusion

- "Model free" methods for recovering 3D human pose from monocular silhouettes
- Direct nonlinear regression of pose again robust shape descriptors
- Tested different regression methods: ridge regression, RVM, SVM
- Pose recovery from static images and image sequences

Work supported by European projects VIBES, LAVA and PASCAL